The Community Foehn Classification Experiment

Georg J. Mayr, David Plavcan, Laurence Armi, Andrew Elvidge, Branko Grisogono, Kristian Horvath, Peter Jackson, Alfred Neururer, Petra Seibert, James W. Steenburgh, Ivana Stiperski, Andrew Sturman, Željko Večenaj, Johannes Vergeiner, Simon Vosper, and Günther Zängl

Many processes and phenomena in the atmosphere need to be diagnosed—from low pressure systems with fronts in midlatitudes and hurricanes in the tropics to fog or lightning. Some diagnoses are easy to make. Hearing thunder identifies lightning, and not being able to see a building less than 1 km away during daytime indicates fog. These diagnoses can even be automated with suitable instrumentation—to identify lightning from its signature in the electromagnetic waves it emits and fog from scattering of a light source. Some processes and phenomena, however, are much harder to classify, often because not enough information is available or the process itself is insufficiently understood. Lately, methods from statistics and machine learning in combination with a huge increase in computing power have been harnessed with ever-increasing success to tackle more and more difficult classification tasks, earning them the label "artificial intelligence." Arguably the greatest progress has been made in classifying images, from spotting a dog in a photo to identifying a particular person. The underlying neural-network algorithms, however, typically need thousands or even hundreds of thousands of preclassified images provided by humans in order to "learn." Such "supervised" learning is much easier than "unsupervised learning" for which no "truth" exists. This is the area where classifications by human experts are still the gold standard, albeit with several drawbacks: lack of scalability and reproducibility, as well as unknown error rates. Because only a few people have the required expertise to perform a classification, which takes a substantial amount of time, the classification task cannot be extended to an arbitrarily large number of instances, and comparisons of classifications among different experts or by the same expert performed at different times are at best extremely rare.

A group of experts collaborated recently on such a task to remedy two of the classification drawbacks by providing estimates of classification uncertainty and reproducibility, and a database against which existing and future algorithms can be tested. The classification task identified periods of downslope windstorms in time series of weather station measurements.

Such windstorms result from winds that cross topographic obstacles and accelerate as they descend to their lee. They occur over mountainous locations worldwide and are known by different names, which are sometimes also used to refer to an additional characteristic. Because no all-encompassing name exists, this article will use "foehn" for simplicity without implying a temperature increase during its onset, or a specific region. Foehn winds affect local weather and climate and impact agriculture (growing conditions due to temperature and humidity changes; top soil erosion), tourism (reliable spots for wind and kite surfing), artificial snow making (change of wet-bulb temperatures), air pollution (trapping pollutants in cold pools underneath the foehn layer, or sweeping

AFFILIATIONS: Mayr, Plavcan,* and Stiperski—University of Innsbruck, Innsbruck, Austria; Armi—Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California; Elvidge—University of East Anglia, Norwich, United Kingdom; Grisogono and Večenaj—University of Zagreb, Zagreb, Croatia; Horvath—Meteorological and Hydrological Service, Zagreb, Croatia; Jackson—University of Northern British Columbia, Prince George, British Columbia, Canada; Neururer and Vergeiner—Central Institution for Meteorology and Geodynamics, Innsbruck, Austria; Seibert—University of Natural Resources and Life Sciences, Vienna, Austria; Steenburgh—University of Utah, Salt Lake City, Utah; Sturman—University of Canterbury, Christchurch, New Zealand; Vosper—Met Office, Exeter, United Kingdom; Zängl—Deutscher Wetterdienst, Offenbach, Germany
* Current affiliation: UBIMET, Vienna, Austria
CORRESPONDING AUTHOR: Georg J. Mayr, georg.mayr@uibk.ac.at

DOI:10.1175/BAMS-D-17-0200.1

them away in case of breakthrough), human health (reduction of air pollution), forest fires (intensifying them to uncontrollable extents), ground traffic (toppling trucks; snow or sand drifts; blasting of vehicles with sand and small rocks), and air traffic (closure of runways when crosswinds are too high). The increasing density of automatic weather stations allows the observation of such winds at progressively more locations. Classification, however, is difficult because other wind systems, such as radiatively driven downslope/downvalley winds, might be superimposed on foehn or share some of its characteristics, or because not enough information is available. The difficulty is compounded because no unanimously agreed-upon definition of foehn and its indications exist, foehn occurs in a variety of synoptic-scale and mesoscale settings, and different names are being used depending on the region, the sign of the temperature change at its onset, and its depth.

CLASSIFICATION TASK. Nevertheless, two unanimously agreed-upon characteristics are that air crosses an obstacle and that it descends and accelerates on the downwind side causing strong winds. A fairly simple conceptual model of the flow situation after the onset of foehn, corroborated by field campaigns, laboratory experiments, computer simulations, and theoretical investigations, is shown in Fig. 1. Unfortunately, no continuous measurements covering the vertical cross section are routinely available for classification; only weather stations at the ground provide the necessary observations. Nowadays, with the proliferation of automatic weather stations and mesonets in some regions, measurements close to the crest of the obstacle are also available so that the first foehn characteristic of air *crossing* the topographic obstacle can be checked. The second characteristic, that air *descends*, leads to adiabatic warming and consequentially to a decrease in relative humidity. This pattern can be examined through differences between the crest and a downwind station of variables that are approximately conserved in foehn flow, such as potential temperature or mixing ratio.

Classification is made more ambiguous by processes for which potential temperature and mixing ratio are not conserved, that is, turbulent mixing within the foehn flow, at the surface and its upper interface; mixing air in from tributaries; phase changes of water (formation and evaporation of liquid and solid particles); and daytime warming and nighttime cooling due to surface sensible heat flux. How large these diabatic effects are varies with the season, time of day, location, and large-scale and mesoscale flow configurations. Information about their contribution is not readily available so that classifications become difficult and possess an unknown and variable degree of uncertainty.

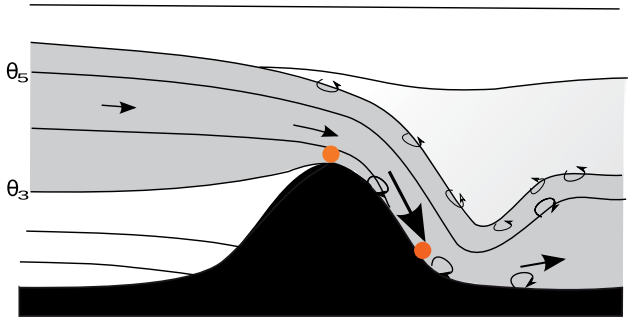

Fig. 1. Conceptual model of well-established foehn flow (dark gray shading) along a vertical cross section exhibiting the two core characteristics of air crossing the obstacle (black), which can be ridge-like, a strait, or a pass, and descending to its lee. Flow is approximately along isentropes (thin lines); straight arrows indicate wind speed, curved ones turbulence. Colored dots are exemplary weather station locations at crest and downwind (cf. Table 1).

THE COMMUNITY FOEHN CLASSIFICATION EXPERIMENT. The Community Foehn Classification Experiment set out to quantify the uncertainty of human foehn classifications, to compare them to machine classifications, and to provide a dataset for the development of foehn classification algorithms. Three groups of human experts and two objective algorithms faced the task of identifying foehn periods. The first group (most of them are coauthors of this paper) consisted of 26 seasoned experts in mountain meteorology from different continents with operational or research backgrounds and thus a broad range of concepts of what constitutes foehn. The other two groups were made up of students taking the advanced weather forecasting course at the University of Innsbruck in 2016 (34) and 2017 (18), respectively. The student groups had a fairly homogeneous level of expertise because they had received four hours of lectures on foehn and had to apply it in homework problems in their advanced weather forecasting course. It was explained to the students why it was crucial for the outcome of this study that they worked completely independently. In addition to

Table 1. Weather station locations used for foehn classification with their long-term foehn frequencies determined from automatic algorithms A1 and A2.

Location	Lat (°N)	Lon (°E)	Alt (m MSL)	Frequency from A1 (%)	From A2 (%)	Town
1	46.30287	7.84294	639	6	10	Visp
2	46.88702	8.62181	438	5	5	Altdorf
3	47.12745	9.51753	457	4	4	Vaduz
4	47.42546	9.39847	776	2	2	St. Gallen
5	47.03643	8.30097	457	< 1	< 1	Luzern
C (crest)	46.65346	8.61625	2287	—	—	Guetsch

human experts, two algorithms were used that also employ the concept shown in Fig. 1. One, labeled A1 henceforth, is in operational use by the Swiss weather service. It uses percentiles of the distribution of the difference of potential temperature between crest and downstream locations (small; cf. Fig. 1), wind speed (high), and relative humidity (low) as hard thresholds for the classification of three categories: no foehn, foehn air mixed with cold valley air, and foehn. The second algorithm, A2, in operational use at the University of Innsbruck, learns from the data by itself and does not use hard thresholds. It uses so-called statistical mixture models to fit two or more parametric distributions to the observed distribution of classifying variables, such as potential temperature difference between crest and downwind stations, and wind speed, to yield a probability for foehn between 0 and 1, instead of merely a binary yes–no classification. Both algorithms require that the appropriate directional sector for foehn winds be manually set.

The classification experiment was designed to strike a balance between ideal goals and practical feasibility for the human classifiers. Therefore, five topographically different locations of differing annual foehn frequencies in the Swiss Alps were selected (Table 1 and Fig. 2). Twelve 48-h periods at each station yielded a total of 60 cases, for which the experts had to classify south foehn periods lasting at least 1 h at 30-min resolution. One of the coauthors, who did not himself manually classify (D. Plavcan), selected these cases based on results from the two automated classification algorithms, A1 and A2, to cover all permutations: phases of foehn–no foehn for which both, only one, or none agreed. Cases contained none, one, or several foehn periods, respectively. Unbeknownst to the classifiers, one difficult 48-h period appeared twice in order to estimate reproducibility.

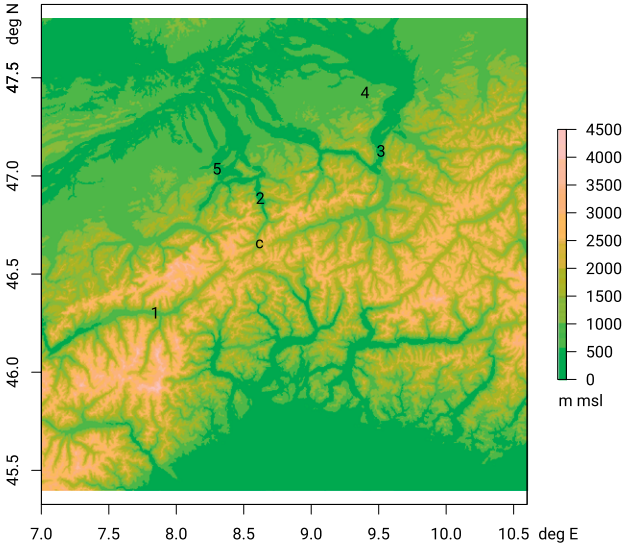

Fig. 2. Topography (m MSL) and location of stations for which foehn was classified (cf. Table 1). Measurements at the crest location (C) were used to assist in the classification at all locations. Digital elevation model at 250-m horizontal resolution from the Shuttle Radar Topography Mission (SRTM; https://cgiarcsi.community/data/srtm-90m-digital-elevation-database-v4-1/).

Each participant received a wind speed–coded wind rose for each location, a pseudo-3D image of the location from Google Earth, exact coordinates, plots of meteorological variables for each of the 60 periods of 48 h, and instructions that contained an annotated example of an additional case reproduced here in Fig. 3. To classify only south foehn events, air had to cross the Alpine crest from south to north as indicated by the wind direction at the crest plotted in black instead of gray, which is fulfilled for the whole 48-h period in this case. Three periods of foehn are

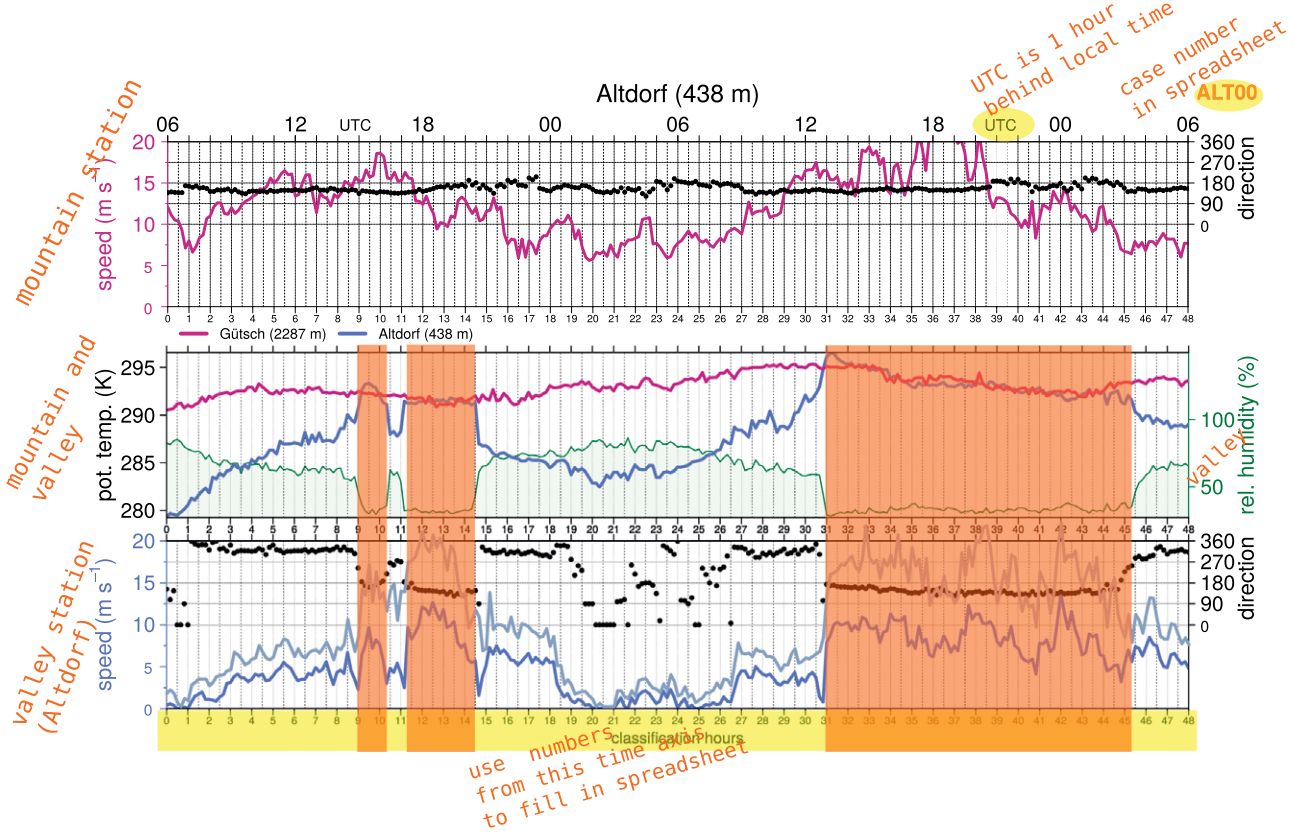

Fig. 3. Annotated time series of an additional case at location 2 (Altdorf) supplied to classifiers with the instruction package and other material. (top) Wind speed (magenta) and direction (° from N; in gray but boldface black when from the foehn sector) at the crest station (2287 m MSL). (middle) Potential temperature at crest station (magenta) and classification location (blue) and relative humidity at classification location (green shaded). (bottom) Wind speed (blue), gusts (light blue), and direction (black; ° from N) at location 2. All values (except gusts) are averages over the previous 10 min. A hypothetical but not unreasonable classification of three foehn episodes at the Altdorf station is marked by orange rectangles (9:00–10:20, 11:10–14:30, and 31:00–45:20). Foehn episodes had to be classified at a resolution of complete half-hour segments and a minimum duration of 1 h. In this example, foehn was classified between 9.0 and 10.0, 11.5 and 14.5, and 31.0 and 45.0.

inferred: from 9:00 to 10:20, 11:10 to 14:30, and 31:00 to 45:20 (as hh:min). During these periods, similar potential temperatures at crest and the classification location imply the second foehn characteristic of lee-slope descent. Wind directions are from the appropriate sector[1] and wind speeds are higher. Temperatures increase at the onset of each period, presumably when foehn erodes an underlying shallow cold pool. Humidity also drops, reflecting the drawdown of drier air from higher altitudes. Because *relative* humidity (%) instead of specific humidity (g kg^{-1}) is plotted, the temperature increase additionally contributes to a drop in relative humidity.

[1] Deduced from wind roses and topography maps; not shown.

RESULTS. The three human groups classified foehn duration during the 12 × 48 h periods at each of the 5 locations broadly similarly, as Fig. 4 shows. Median durations (colored horizontal lines) are within a few percentage points of each other. The group of mountain meteorology experts has the greatest diversity of backgrounds and consequently the most varied concepts of what constitutes foehn. As a result, their classification variation is larger than that of the second group of students, who all had the same foehn concept instilled in their course. The variation of the first group of students, on the other hand, is larger—mainly because of a few outliers at each location.

The variation and thus classification uncertainty is smallest at location 4, a station at the northern edge of the Alps. The largest uncertainty occurred

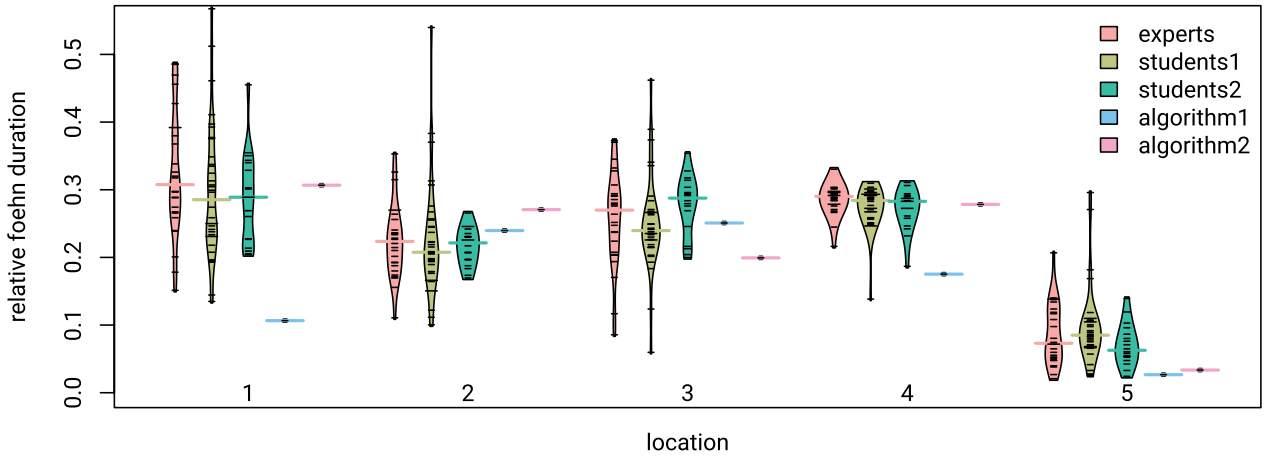

Fig. 4. Beanplots of classified foehn duration at each location relative to the total duration of the time series of 12 x 48 h stratified by classifier groups: experts, two master's student groups, and the two algorithms (foehn "yes" when mixed or pure foehn category are diagnosed in A1, and when foehn probability ≥ 50% is diagnosed in A2). Black lines indicate individual classifications, and colored lines the median of each group. Areas are the empirical densities of each group.

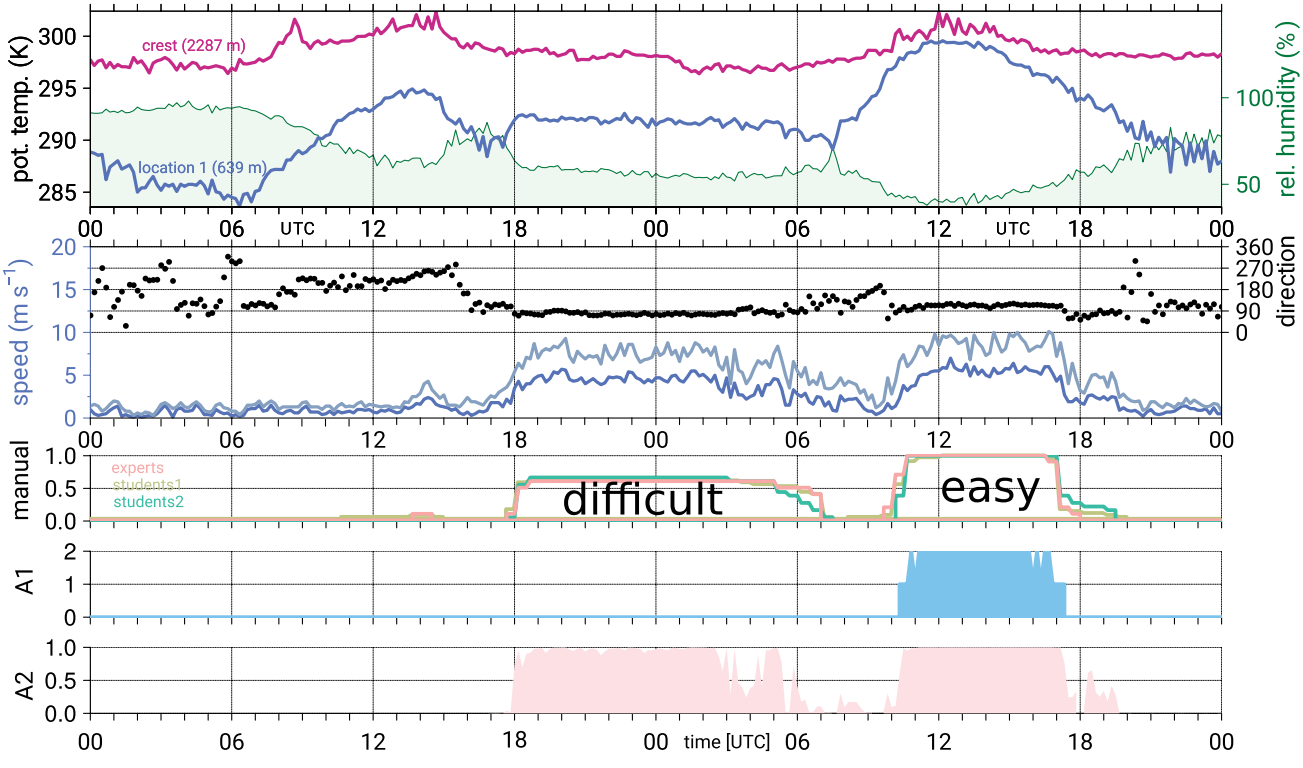

Fig. 5. Classification case at location 1. From top to bottom the panels show the potential temperature (blue) and relative humidity (green) with added potential temperature at crest (purple); the wind direction (black), wind speed average (dark blue), and gusts (light blue) at location 1; the proportion of human classifier groups that classified foehn during the time series; the classifications with the three-category algorithm A1 (no foehn, 0; foehn mixed with valley air, 1; and foehn, 2); and the probability of foehn from the statistical mixture model A2.

for location 1, where foehn can potentially blow from several wind sectors and for which the crest station might not always be representative of the upstream conditions.

The agreement between the algorithms and human classifications varies. Results for A1 are within a few percentage points of the medians of the human groups at locations 2 and 3 and for A2 at locations 1 and 4. However, they are at the margins of human classifications for locations 2 (A2), 3 (A2), and 5 (A1 and A2), and A1 is even outside at locations 1 and 4.

Classification example. Figure 5 shows the classifications from the three groups of human classifiers and the two algorithms for one of the 60 cases. At about midday of the second day, the potential temperature at valley station 1 reached a value close to that of the crest station (purple line), indicating descent of air. Wind speeds also increased. In the evening the signals in the variables reverse, indicating the cessation of foehn conditions. Human classifications agree on a core period of foehn from 11:00 to 14:30 (labeled "easy" in Fig. 5) but differ in onset and end times, with end times less unanimous than onset times. The two algorithms classify similarly.

The nighttime period between days 1 and 2, on the other hand, is more difficult. About 60% of the experts and students classified it as foehn (labeled "difficult"), again agreeing for the core period but differing for onset and even more so for end times. On the evening of the first day the wind direction changed into the foehn sector. At the same time, both average and peak wind speeds increased and potential temperature also increased. Unlike the easy period, however, potential temperature is 5 K colder than at the crest, which likely led the other 40% to classify it as a radiatively cooled nocturnal downslope/downvalley flow. Air originating from a different level than represented by the crest station (cf. Fig. 2) and mixing of foehn air with radiatively cooled air from the valley and its tributaries might have been responsible for such a large difference. The three-category algorithm A1 classifies no foehn, whereas the mixture model algorithm A2 gives a probability close to 1 that it is foehn. The decrease and fluctuations of the probability toward the end of the period stems from the decrease and fluctuations in wind speed and later on the increase in potential temperature difference.

This "difficult" period indicates that a simple yes or no might not be enough for all applications when it comes to classifying foehn flows, for example because of the superposition of foehn and a radiatively cooled downvalley wind. Algorithm A1 adds the third category of "mixed foehn/valley air" (although it does not classify it as such in this particular case). Algorithm A2 gives a continuous probability of foehn occurrence.

Changes in classification uncertainty. Over all 60 cases, delineating the beginning and end of a foehn event had a higher variability among all classifiers. Although the majority of the classified foehn events started with a temperature increase, the uncertainty was not clearly different from the events that started with no change or a decrease in temperature. Classification uncertainty was also higher for the nighttime compared to the daytime for similar reasons as in the difficult period in Fig. 5. Classification uncertainty also varied somewhat seasonally, with low uncertainty in the fall [September–November (SON)] and winter [December–February (DJF)] months; the highest uncertainty in the spring [March–May (MAM)], particularly among human classifiers; and medium uncertainty in the summer months [June–August (JJA)].

Reproducibility. To evaluate reproducibility, one of the more difficult cases (at location 1) occurred twice in the dataset, unbeknown to the classifiers. Figure 6 shows the relative frequency of the absolute difference of the foehn duration classified at the first occurrence and the second occurrence of that case. Ideally and for perfect reproducibility, the difference in classified foehn duration among the identical cases is zero. However, fewer than half of the classifiers achieved perfect reproducibility.

This lack of reproducibility is worrisome, although probably less extreme for easier cases. Nevertheless, it

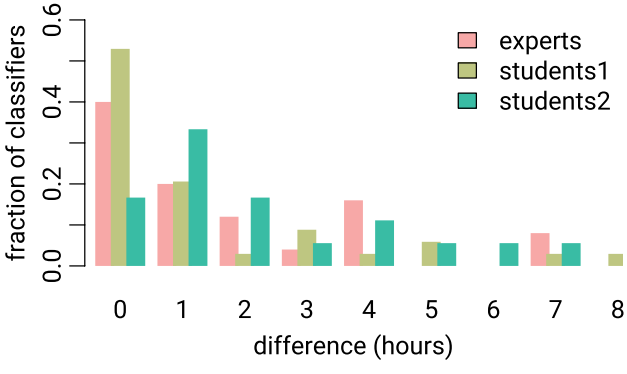

Fig. 6. Histogram of the absolute difference in classified foehn duration (h) between two identical cases. For perfect reproducibility, all classifiers should have had 0-h difference. The bars are for the hour prior to and including the labeled duration difference.

corroborates the first author's personal experience of classifying foehn at different locations globally.

Dataset. The dataset will be available from the University of California (UC), Irvine, which hosts a large repository of classification datasets (https://archive.ics.uci.edu/ml/about.html).

CONCLUSIONS. Several lessons have been learned from this experiment that add on the one hand supporting evidence to what was previously at least informally known from other classification tasks (points i–iii below), and on the other hand (points iv–vi) add new knowledge. i) Busy experts are willing to volunteer a chunk of their scarce time provided the classification task is an intellectually challenging puzzle. ii) Human experts use implicit (and in the case of the master's students, explicitly taught) physically based concepts to help them distinguish between the two categories of foehn–no foehn. iii) Expert classifications carry uncertainty and are not even necessarily reproducible, which needs to be quantified (as here) or at least considered when interpreting results using such classifications. iv) Uncertainty is largest for onset and even more so for the ending of a foehn event and also larger during the night. v) Combining advanced statistical and/or machine learning models with physically based concepts for choosing their input variables yields similar results to those of human experts. In addition, they easily scale to longer time series or more locations and are reproducible, which is a fundamental scientific requirement and allows the comparison of different datasets (foehn occurrence at different locations in this case). It is thus highly recommended to develop objective classification procedures, ideally without having to resort to manually specified and/or hard limits. If the algorithms are additionally made available as packages of open-source languages, foehn classifications can easily be reproduced by other researchers. vi) Diagnoses contain more information when they are probabilistic instead of binary yes–no—a concept that has a long history of implementation in (weather) forecasts.

In addition to shedding light on human and machine classification of foehn, the dataset allows the testing of existing and newly developed algorithms for unsupervised learning tasks when truth is not known, such as in the case of foehn occurrence. It can also serve a community interested in estimating the accuracy of previous human foehn classifications and climatologies.

ACKNOWLEDGMENTS. The authors give many thanks to Achim Zeileis for discussions on how to best design the experiment, and a huge thanks to the experts and the 2016 and 2017 cohorts of the Advanced Weather Forecasting class who classified these 60 periods.

FOR FURTHER READING

Brinkmann, W. A. R., 1971: What is a foehn? *Weather,* **26,** 230–240, https://doi.org/10.1002/j.1477-8696.1971.tb04200.x.

Dürr, B., 2008: Automatisiertes Verfahren zur Bestimmung von Föhn in Alpentälern. Arbeitsberichte der MeteoSchweiz Tech. Rep. 223, 22 pp., www.meteoschweiz.admin.ch/content/dam/meteoswiss/de/Ungebundene-Seiten/Publikationen/Fachberichte/doc/ab223.pdf.

Elvidge, A. D., and I. A. Renfrew, 2016: The causes of foehn warming in the lee of mountains. *Bull. Amer. Meteor. Soc.,* **97,** 455–466, https://doi.org/10.1175/BAMS-D-14-00194.1.

Jackson, P. L., G. Mayr, and S. Vosper, 2013: Dynamically-driven winds. *Mountain Weather Research and Forecasting: Recent Progress and Current Challenges,* F. K. Chow, S. F. J. de Wekker, and B. J. Snyder, Eds., Springer, 121–218, https://doi.org/10.1007/978-94-007-4098-3_3.

Leisch, F., 2004: FlexMix: A general framework for finite mixture models and latent class regression in R. *J. Stat. Softw.,* **11,** 1–18, https://doi.org/10.18637/jss.v011.i08.

Plavcan, D., G. J. Mayr, and A. Zeileis, 2014: Automatic and probabilistic foehn diagnosis with a statistical mixture model. *J. Appl. Meteor. Climatol.,* **53,** 652–659, https://doi.org/10.1175/JAMC-D-13-0267.1.

Richner, H., and P. Hächler, 2013: Understanding and forecasting Alpine foehn. *Mountain Weather Research and Forecasting: Recent Progress and Current Challenges,* F. K. Chow, S. F. J. de Wekker, and B. J. Snyder, Eds., Springer, 219–260, https://doi.org/10.1007/978-94-007-4098-3_4.

Seibert, P., 2012: The riddles of foehn—Introduction to the historic articles by Hann and Ficker. *Meteor. Z.,* **21,** 607–614, https://doi.org/10.1127/0941-2948/2012/0398.

Von Ficker, H., 1910: On the formation of foehn winds on the northern side of the Alps. *Meteor. Z.,* **21,** 597–605, https://doi.org/10.1127/0941-2948/2012/0532.

ams EDUCATION

THANK YOU

- MORE THAN 21,000 TEACHERS DIRECTLY TAUGHT
- OVER 175,000 TEACHERS PEER TRAINED
- MILLIONS OF STUDENTS BENEFIT

ametsoc.org/educationprogram

Thank you for supporting the AMS Education Program! Your support contributes to our ability to provide professional development programs and opportunities in weather, ocean and climate science to K-12 teachers.